Peningkatan Performa Algoritma Naive Bayes dengan Gain Ratio untuk Klasifikasi Keganasan Kanker Payudara

Muhammad Faizal Kurniawan, Jusak Nugraha Irawan
STMIK Widya Pratama Pekalongan
E-mail: m.faizalkurniawan@gmail.com, jusak_n129@gmail.com

RINGKASAN

Kanker adalah salah satu penyakit yang sampai saat ini memakan banyak korban jiwa. Tercatat dalam 5 tahun dari data tahun 2012 International Agency for Research of Cancer (IARC) merilis ada lebih dari 14 juta jiwa mengidap penyakit kanker dan 8,2 juta diantaranya meninggal dunia karena kanker yang diderita. Dari data tersebut jenis kanker yang paling banyak diderita adalah kanker payudara yaitu 19,2% dari keseluruhan 14 juta kasus lebih. Pencatatan terkait pasien dan jenis kanker banyak dilakukan di dunia medis. Data tersebut semakin banyak dan hanya akan menjadi sampah apabila tidak dapat digunakan sebagai pengetahuan baru. Data mining merupakan bidang ilmu yang menjawab tantangan banyaknya data. Klasifikasi merupakan bagian dari data mining yang memungkinkan penciptaan informasi dan pengetahuan baru dari data lampau. Salah satu teknik klasifikasi terbaik dan terbukti banyak digunakan adalah naive bayes. Dari penelitian thun 2016 naive bayes memiliki performa yang terbaik untuk klasifikasi penyakit kanker payudara. Dataset yang besar dengan atribut yang banyak tidak menjamin performa algoritma akan lebih baik. Salah satu proses peningkatan performa algoritma adalah dengan melakukan seleksi fitur. Gain ratio merupakan pengembangan dari algoritma information gain yang terbukti handal dan dapat menangani data berdimensi tinggi. Penelitian ini membuktikan bahwa penggunaan algoritma seleksi fitur gain ratio dapat meningkatkan performa naive bayes dalam klasifikasi dataset breast cancer wisconsin. Performa naive bayes tanpa seleksi fitur adalah 92,7% sedangkan setelah dilakukan seleksi fitur menggunakan gain ratio akurasinya naik 4,01% menjadi 96,71%.

Kata Kunci : Data Mining, gain ratio, breast cancer wisconsin, naive bayes

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker merupakan salah satu penyakit berbahaya yang banyak menimbulkan kematian pada penderitanya. Yang terbaru aktris Julia Peres dinyatakan meninggal dunia setelah terdeteksi mengidap kanker stadium akhir. Di seluruh dunia kanker tercatat sebagai penyakit yang mengakibatkan kematian terbesar. Lebih dari 8,2 juta jiwa meninggal dunia akibat penyakit kanker (Iarc. 2012). Berdasarkan data yang didapatkan dari *GLOBOCAN, International Agency for Research of Cancer* (IARC) pada tahun 2012 setidaknya ada 14.067.894 penderita kanker baru dan menyebabkan kematian atas 8.201.575 jiwa. Dari banyaknya jenis kanker, kanker payudara merupakan jenis kanker yang terbanyak diderita di seluruh dunia dalam 5 tahun terakhir.

Tabel 1 menrepresentasikan data penderita penyakit kanker dalam 5 tahun terakhir (Iarc. 2012). Dalam tabel tersebut *Breast cancer* atau kanker payudara merupakan jenis kanker dengan nilai presentase terbesar dengan jumlah penderita sebanyak 19,2% dari keseluruhan penderita kanker. Banyak penderita penyakit kanker yang mengetahui diagnosa setelah mengalami beberapa komplikasi. Ketika penyakit kanker terdeteksi ternyata penyakit tersebut sudah mengalami perkembangan di dalam tubuh. Secara medis penyakit dengan taraf stadium yang lebih tinggi akan lebih sulit ditangani dibandingkan dengan penyakit yang terdeteksi lebih dini.

Perkembangan komputer dan informatika saat ini sungguh sangat pesat. Dalam kondisi sekarang ilmu komputer dan informatika tidak hanya dapat dihubungkan dengan ilmu teknik dan matematis saja. Ilmu komputer dapat diimplementasikan di semua bidang yang ada. Salah satunya adalah di bidang kesehatan. Data mining merupakan satu bidang ilmu yang memanfaatkan data yang sebelumnya kurang terpakai untuk mendapatkan suatu informasi atau pengetahuan baru. Teknik data mining dapat digunakan untuk proses klasifikasi dengan memanfaatkan data lampau.

Tipe data sangat mempengaruhi performa dan akurasi suatu algoritma (Amancio et al. 2013)

Tabel 1. Data Penderita Kanker (Iarc. 2012)

Cancer	Incidence			Mortality			5-year prevalence		
	Number	(%)	ASR (W)	Number	(%)	ASR (W)	Number	(%)	Prop
Lip, oral cavity	300373	2.1	4	145353	1.8	1.9	702149	2.2	13.5
Nasopharynx	86691	0.6	1.2	50831	0.6	0.7	228698	0.7	4.4
Other pharynx	142387	1	1.9	96105	1.2	1.3	309991	1	6
Oesophagus	455784	3.2	5.9	400169	4.9	5	464063	1.4	8.9
Stomach	951594	6.8	12.1	723073	8.8	8.9	1538127	4.7	29.6
Colorectum	1360602	9.7	17.2	693933	8.5	8.4	3543582	11	68.2
Liver	782451	5.6	10.1	745533	9.1	9.5	633170	2	12.2
Gallbladder	178101	1.3	2.2	142823	1.7	1.7	205646	0.6	4
Pancreas	337872	2.4	4.2	330391	4	4.1	211544	0.7	4.1
Larynx	156877	1.1	2.1	83376	1	1.1	441675	1.4	8.5
Lung	1824701	13	23.1	1589925	19	19.7	1893078	5.8	36.5
Melanoma of skin	232130	1.7	3	55488	0.7	0.7	869754	2.7	16.8
Kaposi sarcoma	44247	0.3	0.6	26974	0.3	0.3	80395	0.2	1.5
Breast	**1671149**	**12**	**43.1**	**521907**	**6.4**	**12.9**	**6232108**	**19**	**240**
Cervix uteri	527624	3.8	14	265672	3.2	6.8	1547161	4.8	59.6
Corpus uteri	319605	2.3	8.3	76160	0.9	1.8	1216504	3.7	46.8
Ovary	238719	1.7	6.1	151917	1.9	3.8	586624	1.8	22.6
Prostate	1094916	7.8	30.7	307481	3.7	7.8	3857500	12	149
Testis	55266	0.4	1.5	10351	0.1	0.3	214666	0.7	8.3
Kidney	337860	2.4	4.4	143406	1.7	1.8	906746	2.8	17.5
Bladder	429793	3.1	5.3	165084	2	1.9	1319749	4.1	25.4
Brain, nervous system	256213	1.8	3.4	189382	2.3	2.5	342914	1.1	6.6
Thyroid	298102	2.1	4	39771	0.5	0.5	1206075	3.7	23.2
Hodgkin lymphoma	65950	0.5	0.9	25469	0.3	0.3	188538	0.6	3.6
Non-Hodgkin lymphoma	385741	2.7	5.1	199670	2.4	2.5	832843	2.6	16
Multiple myeloma	114251	0.8	1.5	80019	1	1	229468	0.7	4.4
Leukaemia	351965	2.5	4.7	265471	3.2	3.4	500934	1.5	9.6
All cancers excl. non-melanoma skin cancer	14067894	100	182	8201575	100	102	32455179	100	625

Data mining melakukan perhitungan matematis dan algoritmik untuk mendapatkan pengetahuan dari sebuah data. Algoritma data mining terbaik untuk satu tipe data belum tentu baik untuk tipe data yang lain (Patel, Vala, and Pandya 2014). Algoritma terbaik dalam komparasi data mining dapat saja menjadi buruk apabila data yang digunakan memiliki karakteristik yang berbeda (Ragab et al. 2014) (Ashari, Paryudi, and Tjoa 2013). Salah satu algoritma data mining yang terbaik dan banyak digunakan untuk klasifikasi dataset nominal adalah *Naive Bayes* (Wu et al. 2007). Penelitian sebelumnya melakukan komparasi antara beberapa algoritma data mining guna mengetahui algoritma terbaik untuk dataset kanker payudara. Algoritma Naive Bayes

merupakan algoritma terbaik untuk klasifikasi penyakit kanker payudara (Kurniawan and Ivandari 2017).

Dalam klasifikasi data mining memungkinkan perhitungan setiap atribut data yang ada. Banyaknya atribut dapat mempengaruhi performa sebuah algoritma (Maimoon 2010). Atribut yang relevan akan meningkatkan performa algoritma dan sebaliknya, banyaknya atribut yang tidak relevan akan membuat performa algoritma menjadi kurang baik (Han and Kamber 2006a).Tipe dari attribut dataset juga sangat mempengaruhi performa suatu algoritma (Alpaydin 2010). Salah satu metode untuk mencari atribut yang relevan adalah dengan menghitung seluruh kepentingan atribut terhadap dataset yang ada. Setelah seluruh atribut diketahui bobot kepentingannya proses berikutnya yang dapat dilakukan adalah menghilangkan atribut yang tidak relevan. Proses pemotongan atribut ini disebut juga dengan istilah seleksi fitur. *Gain ratio* merupakan algoritma seleksi fitur yang baik dan dapat menangani atribut yang banyak. *Gain ratio* adalah pengemnagan dari algoritma *information gain* (Koprinska 2010).

Penelitian ini akan melakukan perhitungan seleksi fitur dataset *breast cancer wisconsin* dengan *gain ratio* untuk meningkatkan performa algoritma klasifikasi naive bayes.

2. METODE PENELITIAN

Penelitian ini adalah penelitian eksperimental yang melakukan pengujian terhadap hasil beberapa pengukuran akurasi dari algoritma naïve bayes. Data yang digunakan dalam penelitian ini adalah dataset *breast cancer wisconsin* yang merupakan dataset *public* dari UCI *repository*. UCI *repository* merupakan salah satu sumber dataset terpercaya yang menyediakan lebih dari 347 dataset *machine learning*. Dataset dari UCI banyak digunakan oleh peneliti bidang ilmu komputer untuk menguji metode atau model suatu algoritma. Dataset *breast cancer wisconsin* secara lebih terperinci dapat diunduh di alamat website: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

:

2.1 Pengumpulan Data

Metode pengumpulan data merupakan tahapan pertama yang dilakukan dalam penelitian ini. Dalam metode pengumpulan data akan digunakan data public yaitu data breast cancer wisconsin yang didapatkan dari UCI repository seperti yang telah dijelaskan sebelumnya. Dataset tersebut memiliki 699 record, 1 atribut id, 9 atribut informasi dan 1 atribut kelas. Dalam dataset tersebut terindikasi ada 458 terjangkit kanker jinak dan 241 lainnya terjangkit kanker ganas. Atribut informasi yang digunakan antara lain: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape,Marginal Adhesion, Single Ephitelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, serta Mitoses.

Tabel 2 merupakan metadata dari dataset yang digunakan dalam penelitian ini, sedangkan dataset secara keseluruhan disampaikan dalam lampiran dan dengan file excell yang akan diunggah dalam catatan harian di portal simlitabmas. Dalam metadata ini terdapat 11 atribut dengan 1 atribut label dan 1 atribut id. Sehingga atribut reguler yang digunakan dalam proses klasifikasi hanya tersisa 9 atribut.

Tabel 2 Metadata *breast cancer wisconsin*

Role	Name	Type	Statistics	Range	Missings
id	Sample code number	numeric	avg = 1071704.099 +/- 617095.730	[61634.000 ; 13454352.000]	0
label	Class	numeric	avg = 2.690 +/- 0.951	[2.000 ; 4.000]	0
regular	Clump Thickness	numeric	avg = 4.418 +/- 2.816	[1.000 ; 10.000]	0
regular	Uniformity of Cell Size	numeric	avg = 3.134 +/- 3.051	[1.000 ; 10.000]	0
regular	Uniformity of Cell Shape	numeric	avg = 3.207 +/- 2.972	[1.000 ; 10.000]	0
regular	Marginal Adhesion	numeric	avg = 2.807 +/- 2.855	[1.000 ; 10.000]	0

regular	Single Epithelial Cell Size	numeric	avg = 3.216 +/- 2.214	[1.000 ; 10.000]	0
regular	Bare Nuclei	numeric	avg = 3.545 +/- 3.644	[1.000 ; 10.000]	16
regular	Bland Chromatin	numeric	avg = 3.438 +/- 2.438	[1.000 ; 10.000]	0
regular	Normal Nucleoli	numeric	avg = 2.867 +/- 3.054	[1.000 ; 10.000]	0
regular	Mitoses	numeric	avg = 1.589 +/- 1.715	[1.000 ; 10.000]	0

2.2 Desain Eksperimen Algoritma

Tahapan selanjutnya setelah pengumpulan data adalah desain eksperimental algoritma dilanjutkan dengan pengujian algoritma. Tahapan pertama adalah proses seleksi fitur menggunakan algoritma *gain ratio* untuk mendapatkan atribut yang benar benar berpengaruh dalam klasifikasi. Dalam tahap eksperimen digunakan algoritma naive bayes untuk klasifikasi tingkat keganasan kanker payudara. Selanjutnya akan dilakukan validasi dan evaluasi algoritma untuk klasifikasi deteksi penyakit kanker payudara

2.2.1 Seleksi fitur *gain ratio*

Gain ratio merupakan algoritma seleksi fitur yang banyak digunakan peneliti karana handal dan mampu berjalan pada dimensi data yang tinggi. Tahapan seleksi fitur sebenarnya adalah menghitung kepentingan dari keseluruhan atribut data yang ada untuk nantinya dijadikan patokan dalam tahap berikutnya yaitu klasifikasi. Hasil akhir dari tahapan ini adalah atribut yang memiliki tingkat kepentingan yang tinggi selanjutnya akan digunakan sedangkan atribut dengan tingkat kepentingan yang rendah tidak akan digunakan dalam tahap berikutnya.

2.2.2 Tahap Eksperimen

Tahap eksperimen dilakukan dengan menggunakan *tools software* Rapid Miner. Sebelumnya dataset yang sudah dikumpulkan dimasukkan ke dalam aplikasi. Kemudian dilakukan perhitungan dengan menggunakan algoritma *gain ratio* untuk mengetahui tingkat kepentingan atribut. Selanjutnya dilakukan klasifikasi hanya menggunakan atribut terpilih. Proses ini dilakukan secara berulang sampai dengan menghasilkan tingkat akurasi yang terbaik. Dalam tahapan ini digunakan aplikasi bantu rapid miner untuk melakukan perhitungan seluruh algoritma. Hasil akhir tahapan ini adalah prosentase tingkat akurasi dari klasifikasi

2.2.3 Validasi

Dalam proses validasi penelitian ini akan digunakan *10 folds cross validation*. Proses ini banyak digunakan oleh peneliti karena sudah terbukti baik dan menghasilkan tingkat akurasi yang stabil. Secara tori *10 folds cross validation* sudah dijelaskan secara lebih terinci dalam bab sebelumnya (Witten, Frank, and Hall 2011).

2.2.4 Pengukuran akurasi algoritma

Pengukuran dari suatu algoritma merupakan suatu pembuktian yang banyak dilakukan peneliti (Amancio et al. 2013). Dalam prosesnya banyak cara dapat digunakan untuk mengetahui performa suatu algoritma. Salah satu yang paling banyak digunakan adalah dengan menghitung akurasi algoritma. Perhitungan akurasi adalah presentase dari jumlah data *testing* dengan klasifikasi yang sesuai dengan aslinya dibagi keseluruhan data. Cara lain adalah dengan menghitung *Error rate. Error rate* adalah kebalikan dari tingkat akurasi, yaitu presentase kesalahan klasifikasi dibagi dengan keseluruhan dataset.

Dalam penelitian ini digunakan *confusion matrix* sebagai alat ukur performa algoritma klasifikasi. *Confusion matrix* merupakan salah satu alat untuk menghitung nilai akurasi suatu algoritma. Dalam matrix ini dapat dilihat keseluruhan data *testing* yang sesuai dengan klasifikasi sebenarnya serta yang tidak sesuai. Perhitungan secara lebih terperinci dapat dilihat dalam bab sebelumnya.

2.3 Perancangan Sistem Pendukung Keputusan

Tahapan ini adalah tahapan dimana akan dirancang sebuah sistem berdasarkan nilai atau hasil eksperimen sebelumnya. Perancangan sistem akan dilakukan dengan menggunakan lembar kerja tampilan dan flow chart. Metode pengembangan sistem *waterfall* digunakan agar seluruh tahapan yang dilakukan terstruktur dan terukur

3. HASIL DAN PEMBAHASAN

3.1 Seleksi Fitur menggunakan *Gain Ratio*

Tahapan seleksi fitur ini menlakukan perhitungan untuk seluruh atribut dataset dengan menggunakan algoritma seleksi fitur *gain ratio.* Proses ini memungkinkan mendapatkan bobot untuk semua atribut dalam proses klasifikasi selanjutnya. Bobot tertinggi adalah 1 (satu) dan bobot terendah adalah 0 (nol). Atribut dengan nilai bobot 1 merupakan atribut dengan kepentingan tertinggi dalam klasifikasi. Sebaliknya atribut dengan nilai bobot 0 merupakan atribut yang kepentingannya sangat kecil dalam proses klasifikasi atau bahkan mungkin tidak memiliki kepentingan dalam proses klasifikasi. Tabel 3 merupakan hasil perhitungan algoritma *gain ratio* untuk pembobotan atribut data *breast cancer wisconsin.*

Tabel 3 Hasil selelsi fitur data *breast cancer wisconsin*

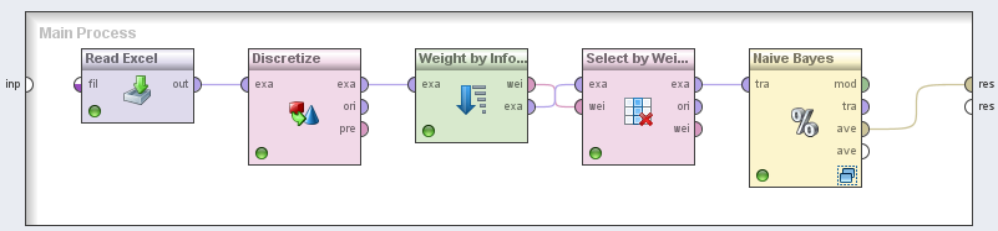
Attribute	***Weight***
Mitoses	0
Marginal Adhesion	0.45203471
Clump Thickness	0.57900365
Bare Nuclei	0.6038166
Single Epithelial Cell Size	0.61652246
Normal Nucleoli	0.62060594
Bland Chromatin	0.75937469
Uniformity of Cell Shape	0.90750813
Uniformity of Cell Size	1

Dari hasil tabel 3 di atas dapat diketahui bahwa atribut *Uniformity of Cell Size* merupakan atribut dengan nilai bobot tertinggi yaitu 1. Artinya atribut *Uniformity of Cell Size* adalah atribut dengan kepentingan tertinggi dibandingkan atribut yang lain. Sedangkan satu atribut lain yaitu *mitoses* merupakan atribut dengan tingkat kepentingan terendah karena memiliki bobot 0 dalam perhitungan menggunakan algoritma seleksi fitur *gain ratio*

3.2 Hasil Akurasi Algoritma Klasifikasi

Setelah diketahui tingkat pembobotan untuk semua atribut yang ada dalam dataset *breast cancer wisconsin* maka dilanjutkan dengan proses klasifikasi menggunakan algoritma naive bayes. Proses ini membandingkan hasil tingkatan akurasi algoritma naive bayes dengan menerapkan ambang batas pada bobot atribut yang ada. Proses ini memungkinkan atribut dengan nilai bobot atribut yang lebih kecil dari nilai ambang batas tidak akan digunakan dalam proses klasifikasi. Dalam hal ini penentuan nilai ambang batas sangat menentukan banyaknya atribut data yang digunakan dalam proses klasifikasi serta pada akhirnya akan mempengaruhi tingkat akurasi dari algoritma tersebut.

Gambar 1 merupakan proses di dalam apliklasi rapid miner yang digunakan dalam penelitian ini:



Gambar 1. Proses penelitian dalam rapid miner

Di dalam gambar tersebut dimulai dari proses paling kiri yaitu membaca dataset yang berasal dari file excell. Dilanjutkan dengan konversi dataset dalam discretizise, lalu dilanjutkan dengan pembobotan menggunakan algoritma seleksi fitur information gain ratio. Proses pembobotan ini menghasilkan luaran yaitu bobot untuk setiap atribut data. Proses selanjutnya yaitu select by weight yaitu melakukan seleksi atribut yang akan digunakan dalam proses selanjutnya dengan memberikan nilai ambang batas sesuai kebutuhan. Proses terakhir yaitu proses klasifikasi algoritma naive bayes. Dalam proses klasifikasi ini dilakikan validasi menggunakan X-Validation dan digunakan 10 folds cross validation karena pembagian iterasi dilakukan selama 10 kali. Serta menggunakan confussion matrix untuk menghitung tingkat akurasi algoritma naive bayes tersebut

Tabel 4 merupakan hasil perhitungan akurasi algoritma naive bayes dengan menggunakan beberapa batas untuk mendapatkan hasil yang terbaik.

Tabel 4 Hasil penelitian

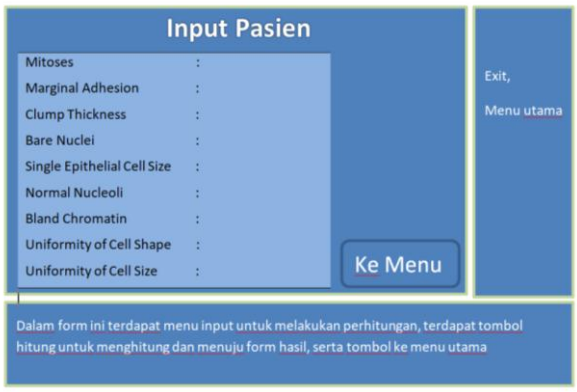
Attribute	Treshold	Accuracy
Mitoses	0	92.7
Marginal Adhesion	0.4	96.42
Clump Thickness	**0.5**	**96.71**
Bare Nuclei	0.6	96.14
Single Epithelial Cell Size	0.61	95.13
Normal Nucleoli	0.62	94.99
Bland Chromatin	0.7	94.7
Uniformity of Cell Shape	0.9	94.42
Uniformity of Cell Size	1	92.7

Tabel 4 menunjukkan tingkat akurasi dari algoritma naive bayes dengan *treshold* yang digunakan untuk seleksi fitur *gain ratio.* Akurasi terbaik didapatkan ketika *treshold* yang digunakan adalah 0.5 dengan tingkat akurasi sebesar 96.71%. Dalam *treshold* ini ada 7 atribut yang digunakan untuk klasifikasi dan hanya menghilangkan 2 (dua) atribut yaitu *mitoses* dan *marginal adhesion.* Kedua atribut tersebut tidak diikutsertakan dalam proses klasifikasi karena keduanya memiliki bobot kurang dari 0.5.

3.3 Perancangan SPK

Dalam tahapan perancangan digunakan Lembar Kerja Tampilan (LKT) untuk mempermudah desain tampilan. LKT yang tercipta sebagaimana gambar 2, 3 dan 4 dibawah ini:

Gambar 2 LKT menu utama



Gambar 3 LKT input

Gambar 4 LKT Hasi Klasifikasi

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Hasil penelitian yang telah dilakukan menemukan bahwa klasifikasi menggunakan algoritma naive bayes untuk dataset breast cancer wisconsin memiliki tingkat akurasi sebesar 92.7%. Hasil ini sudah dianggap baik dan dengan menggunakan keseluruhan atribut data yang ada. Dengan melakukan pre processing yaitu seleksi fitur menggunakan algoritma gain ratio akurasi algoritma naive bayes naik menjadi 96,71%. Hal ini membuktikan bahwa algoritma gain ratio dapat meningkatkan performa dari algoritma naive bayes untuk klasifikasi data breast cancer wisconsin. Kenaikan tingkat akurasi yang didapatkan adalah 4,1%. .

4.2 Saran

Penelitian ini menggunakan aplikasi rapid miner untuk melakukan pembuktian dan perhitungan. Dalam penelitian selanjutnya dapat digunakan pembuktian dengan aplikasi lain seperti halnya aplikasi yang dibuat khusus untuk klasifikasi ini. Atau dapat juga dengan melakukan perhitungan menggunakan microsoft excell

5. UCAPAN TERIMAKASIH

Penelitian ini didanai oleh DRPM Dikti dari hibah Penelitian Dosen Pemula tahun anggaran 2018.

6. DAFTAR PUSTAKA

Alkaromi, M Adib. 2014. "Information Gain Untuk Pemilihan Fitur Pada Klasifikasi Heregistrasi Calon Mahasiswa Dengan Menggunakan K-NN."

Amancio, D. R., C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. a. Rodrigues, and L. Da F. Costa. 2013. "A Systematic Comparison of Supervised Classifiers," October. http://arxiv.org/abs/1311.0202v1.

Ashari, Ahmad, Iman Paryudi, and A Min Tjoa. 2013. "Performance Comparison between Naïve Bayes , Decision Tree and K-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool" 4 (11): 33–39.

Han, Jiawei, and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques Second Edition*. *Elsevier*. Elsevier.

Kusrini, and Luthfi Emha Taufiq. 2009. *Algoritma Data Mining*. Yogyakarta: Andi Offset.

Larose, Daniel T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons.

Patel, Kanu, Jay Vala, and Jaymit Pandya. 2014. "Comparison of Various Classification Algorithms on Iris Datasets Using WEKA" 1 (1): 1–7.

Ragab, Abdul Hamid M., Amin Y. Noaman, Abdullah S. Al-Ghamdi, and Ayman I. Madbouly. 2014. "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining." *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments - IDEE '14*. New York, New York, USA: ACM Press, 106–13. doi:10.1145/2643604.2643631.

Witten, Ian H, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Elsevier.

Wu, Xindong. 2009. *The Top Ten Algorithms in Data Mining*. Edited by Vipin Kumar. New York: Taylor & Francis Group, LLC.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2007. *Top 10 Algorithms in Data Mining*. *Knowledge and Information Systems*. Vol. 14. doi:10.1007/s10115-007-0114-2.